

基于Word2Vec及大众健康信息源的疾病关联探测

罗文馨　陈　翀　邓思艺
(北京师范大学政府管理学院　北京 100875)

摘要: 【目的】利用Word2Vec深度学习技术从面向大众的健康信息中寻找疾病关联, 解决非医学人士通常不了解多种疾病之间存在的关联, 从而影响到健康信息搜寻中的全面性和有效性的问题。【方法】由专家选取30个常见疾病主题, 从高质量医学新闻网站上采集对应疾病的文档, 运用Word2Vec技术对各疾病的相关文档构造词向量, 计算向量距离判断疾病关联。通过与专家评分的相关分析衡量判断结果的准确性。【结果】最优情况下, Word2Vec得到的结果与专家评分相关系数达到0.635。通过对比不同的算法模型、优化方法、数据规模及重要参数对结果的影响, 发现Skip-Gram模型结合负样本数为20的Negative Sampling优化方法在大规模数据集上的实验结果最优。【局限】疾病主题选取宽泛时, 影响Word2Vec判断准确性, 本文的疾病主题选取粒度有待改善。【结论】利用Word2Vec技术在面向大众的健康信息源中也可以探测疾病关联, 其有效性表明该技术可用于改善大众的健康信息搜寻的个性化服务。

关键词: Word2Vec　疾病关联　非专业医学文本　健康信息　个性化
分类号: TP391　G350

1 引　言

以往, 普通大众多从专业医学人士处获取疾病健康知识。互联网的发展可以让大众更加主动地去上网搜寻自己所需的健康信息。近年来, 各种新型健康服务平台不断兴起, 这些服务多以疾病知识科普、在线咨询为主, 极大丰富了人们获取医学信息的渠道。然而, 大众由于缺乏专门的医学知识, 并不了解疾病之间复杂的关联, 例如牙周疾病可能由糖尿病引起。对这种关联缺乏了解会影响到大众管理自身健康、搜寻全面有效的医学信息。如果能通过技术手段寻找疾病主题之间的关联, 可用于改善健康信息的个性化服务, 提高信息服务平台的内容组织和导航质量。由于专业医学文献使用的术语不易被大众理解, 本文使用非专业医学信息, 如高质量的健康新闻, 通过Word2Vec深度学习技术, 基于疾病相关文档探测疾病主题之间的关联, 并与专家评判结果对比, 发现这种技术能有效地用于疾病之间的关联探测。

2 相关工作

面向普通大众的健康信息服务早就引起关注[1], Eysenbach明确提出了结合信息技术手段为消费者提供健康信息服务, 包括分析消费者的健康信息需求, 研究并实现能为消费者提供信息的方法, 依据消费者的偏好设计模型构建信息系统等[2]。国内称这一研究范畴为“用户健康信息学”。目前面向消费者的健康服务不断涌现, 提供疾病知识科普、定制的信息推送或疾病问题在线咨询等, 推动人们管理自身健康, 提高

通讯作者: 陈翀, ORCID: 0000-0002-9704-1575, E-mail: chenchong@bnu.edu.cn。

大众健康信息素养。

为帮助人们更高效准确地获取健康信息，研究人员开展了很多工作，主要分为几个方面:

(1) 调查消费者的信息查寻行为[3]，弄清他们在互联网上查找医学健康信息时最关心什么类型的内容;

(2) 帮助人们理解医学术语，解决由于“词汇之间的鸿沟”带来的难以理解信息或者理解有偏差的问题[4]，例如研制用户健康词表(CHV)[5]、预测用户对健康术语的熟悉度[6];

(3) 建立从医学专业领域概念到普通认知范畴的映射[7]，处理用户健康词汇与 UMLS 词表匹配的问题[8]。

然而，由于疾病之间存在着复杂的关联，未经专业医学训练的普通大众很难了解疾病之间的关联。这影响他们在信息搜寻的时候获取全面的相关信息。目前这方面的研究还比较欠缺。

传统上疾病关联探测是临床医学研究或生物医学实验的任务。现有的利用文本挖掘探测疾病关联的研究主要以专业医学文献为研究对象。比如有学者采用语义扩展模型和神经网络聚类方法，将疾病类型与致病基因关联起来[9-11]。这些研究结论多为分子生物学、基因、化学成分等层面的解释，缺乏专业知识的普通大众是很难理解的。

面向大众的医学健康信息源包括健康门户网站、医学新闻网站、在线健康社区、公共健康知识库等。对在线健康社区 MedHelp 的用户发帖的研究，发现药物与其不良反应的关系，有助于药品安全监管者有效地识别早期药品不良反应信号[12]。对特定疾病社区中的用户帖子进行文本聚类分析，分析三类疾病之间的联系与差异[13]。这些研究说明利用大众健康信息源可以找到一些对用户很有参考意义的联系。但在线健康社区的信息内容质量不佳，为了保证研究结论的可靠性，本文选择高质量的医学新闻。此外，还有利用社会网络分析法来探究健康主题之间关系的研究，如刘红霞等[14]对 WHO 网站的健康信息主题进行分析，采用文本相似性算法，挖掘它们之间的链接关系和语义关系，用社会网络的方式来呈现。但该方法过于依赖特定网站的链接结构，所能找到的关联比较受限；研究中采用文本相似性算法，也没有充分反映其语义层面的关系。

据笔者调研所知，利用大众健康信息挖掘不同疾病或主题间关系的研究有很值得深入的空间。本文将疾病关联的发现任务转换为探测疾病相关文档的语义关联，利用 Word2Vec(Word to Vector)技术找到与特定疾病关系密切的词汇，利用这一桥梁发现疾病关联。

2003 年，Bengio 等提出神经网络语言模型(Neural Network Language Model, NNLM)，利用神经网络结构对自然语言建模的同时，得到了词向量[15]。2013 年，Mikolov 等简化 NNLM 模型，提出 CBOW(Continuous Bag-Of-Words)模型和 Skip-Gram 模型[16]，旨在更高效地实现词语的向量表示。同年，Google 公司推出这两个模型的 C 语言实现版本，称之为 Word2Vec；目前 Python 库中 gensim 包也集成了该算法。Word2Vec 是基于深度学习思想[17]，通过训练文本数据集，将词语不同的语法和句法特征映射到向量的不同维度上去，将单个词语表示为高维向量空间中的某个点。它用于实现词语的向量表示时主要有 CBOW 和 Skip-Gram 两种模型。两者的区别在于，CBOW 模型是已知上下文，预测中心词；而 Skip-Gram 模型则是已知当前词，预测其上下文。相关研究证明，该技术应用在词语相似度计算[18]、机器翻译、特征抽取[19-20]、情感分类[21]等领域效果较好。Word2Vec 技术具有通用性并且使用方法相对较简单。

3 疾病关联探测

不同于以往从生物实验和临床角度寻找疾病关联，本文将探测疾病关联的任务转换为从疾病相关文档中发现语义关联。具体采用 Word2Vec 技术，利用医学健康新闻寻找疾病关联，旨在探讨一种通用的方法找到疾病主题之间的关联关系，改进人们搜寻健康信息的效率和效果。本文主要围绕以下两个问题:

(1) 如何利用Word2Vec寻找疾病之间的关联关系?

(2) 如何评估Word2Vec应用在疾病关联探测上的效果?

在特定疾病的相关文档集合上，用 Word2Vec 技术找到揭示不同疾病主题的词汇向量，借助其相似度确定疾病主题的关联；通过统计分析方法将结果与专家评分结果进行对比，结合参数调优实验确定可令结果最优的参数配置。

3.1 数据采集

不同于以往在专业医学文献中挖掘疾病关联的研

究，本文选用普通人能理解的健康信息，原因是专业医学文档的术语难以为大众所理解，即使找到了关联，也难以直接应用于普通大众经常浏览的信息源。

数据来自于 Medical News Today 网站。该网站新闻由具有医学背景的专业人员撰写，并由网站人工添加类别标签。内容质量较高且易于被普通人理解。其类别标签按大众关心的健康问题分为 144 个类，每个类都有对应的新闻文档。

本研究采用其中 30 种有代表性的疾病类别，Addiction (成瘾)、Allergy (过敏)、Alternative Medicine (补充和代替医疗)、Anxiety (焦虑)、Arthritis (关节炎)、Asthma (哮喘)、Breast Cancer (乳腺癌)、Cardiovascular (心血管)、Cholesterol (胆固醇)、COPD(慢性阻塞性肺疾病)、Dentistry (牙科)、Depression (抑郁)、Diabetes (糖尿病)、Eating Disorders (饮食失调)、Flu (流感)、Headache (头痛)、Heart Disease (心脏病)、HIV (艾滋病)、Hypertension (高血压)、Men's Health (男性健康)、Mental Health (心理健康)、Neurology (神经病学)、Nutrition (营养学)、Obesity (肥胖)、Pregnancy (怀孕)、Prostate (前列腺)、Seniors (老年人疾病)、Sleep (睡眠问题)、Women's Health (女性健康)、Stroke (中风)。

采集每个选定类别的疾病中的健康新闻。对医学健康新闻网页使用 Python 的自然语言工具包 NLTK3.2 版本进行文本预处理，经过清除网页噪音、分词、统一大小写、词形归并、去除停用词等步骤。

为了对比数据集对算法结果的影响，使用的数据集分为 3 000、6 000 和 9 000 个网页三种，分别记为 3K、6K、9K。其中 6K 数据集是在每个类别已经抓取前 100 个网页基础上，又继续抓取 100 个网页得到的，9K 同理。

3.2 Word2Vec 模型构建

Word2Vec[16-17] 用于实现词语的向量表示有 CBOW 模型和 Skip-Gram 模型；用于优化算法效率的方法包括 Hierarchy SoftMax(HS)和 Negative Sampling (NS)两种。将它们两两组合，得到 4 种训练框架，如表 1 所示：

表 1　训练框架

训练框架	Hierarchy SoftMax	Negative Sampling
CBOW	CBOW&HS	CBOW&NS
Skip-Gram	Skip-Gram&HS	Skip-Gram&NS

(1) CBOW 模型和 Skip-Gram 模型

CBOW 模型与 Skip-Gram 模型实际上是对神经网络语言模型(NNLM)的优化。NNLM 是统计语言模型的一种，工作原理见图 1：

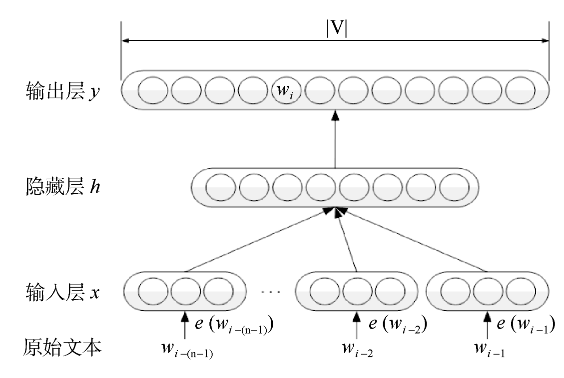


图 1　神经网络语言模型

输入语料库 C，构建词汇表 V，词汇表 V 中词语总量为|V|，假设通过语言模型预测的词为 w_i，其上下文为 w_i 的前(n-1)个词；上述条件下，NNLM 模型目标为最大化式(1):

$$P(w_i \mid w_{i-(n-1)}, w_{i-(n-2)} \cdots w_{i-1}) \quad (1)$$

NNLM 为三层前馈神经网络结构，输入层 x 为前(n-1)个词的词向量的顺序拼接，隐藏层 h，输出层 y 为剩余两层神经网络。其中 H 为输入层到隐藏层的权重矩阵，U 为隐藏层到输出层的权重矩阵，$b^{(1)}$、$b^{(2)}$为偏置项，tanh 为双曲正切函数。

$$x = [e(w_{i-(n-1)}); e(w_{i-(n-2)}) \cdots e(w_{i-1})] \quad (2)$$

$$h = \tanh(b^{(1)} + Hx) \quad (3)$$

$$y = b^{(2)} + Uh \quad (4)$$

值得注意的是，输出层 y 共有|V|个元素，分别对应下一个词为 V 中某词的可能性，需要利用 SoftMax 函数，将其转成概率值:

$$P(w_i \mid w_{i-(n-1)}, w_{i-(n-2)} \cdots w_{i-1}) = \frac{\exp(y(w_i))}{\sum_{j=1}^{N} \exp(y(w_j))} \quad (5)$$

训练时，优化的目标为最大化式(6):

$$\sum_{w_{i-(n-1)}} \log P(w_i \mid w_{i-(n-1)}, w_{i-(n-2)} \cdots w_{i-1}) \quad i \in C \quad (6)$$

在实际训练时，通过随机梯度下降法来不断迭代，每次迭代都会对词向量及训练时中间矩阵等参数进行一次更新。优化完成后，相应的词向量也生成完毕。

由于从隐藏层到输出层的矩阵计算最耗费时间，

故 CBOW 和 Skip-Gram 模型在 NNLM 的基础上去掉了隐藏层，使得计算量大大减小，而准确性则由训练样本的扩大来保证。

CBOW 模型结构见图 2。上下文 c 取词 w_i 的前后各(n-1)/2 个词，假设上下文中所有的词对当前词出现概率影响的权重一样，不考虑出现的先后顺序，将输入层的上下文 c 的词向量 $e(w_i)$拼接改为词向量的平均值(或求和)，如式(7)所示；迭代时优化目标为最大化式(8)，迭代过程中也实现了词向量的优化。

$$x = \frac{1}{n-1}\sum_{w_j \in c} e(w_j) \tag{7}$$

$$\sum_{(w,c)\in C} \log P(w \mid c) \tag{8}$$

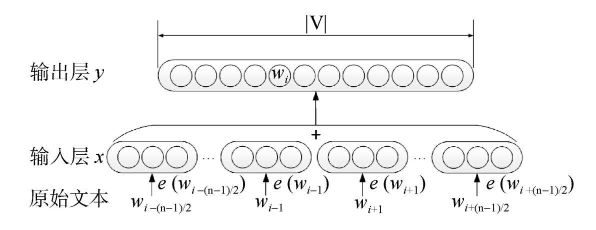


图 2 CBOW 模型

Skip-Gram 模型见图 3，它采用“跳过某些单元”① 的方式来扩大训练样本，上下文词语组合情况增多；从词 w_i 的上下文 c 中随机选择一个词的词向量 w_j 作为输入；优化的目标为最大化式(9):

$$\sum_{(w,c)\in C}\sum_{w_j \in c} \log P(w \mid w_j) \tag{9}$$

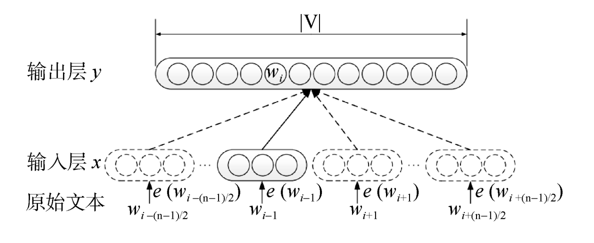


图 3 Skip-Gram 模型

(2) Hierarchy SoftMax 和 Negative Sampling

为了降低模型的时间复杂度，Hierarchy SoftMax 借助分类的方式，对词语按照词频、词性或者主题进行区分，将某个类型下的词群抽象为一个词向量，计算时用这个抽象的词向量代表这类词，从而减小计算的复杂度。比如利用词频特征构造哈夫曼树来进行分层，用抽象的中间节点的向量来近似代替它的所有子节点的向量。Negative Sampling 相对更简单，采用负采样来提高训练速度。模型迭代时，采用随机负采样的方法进行计算并更新，而不是将下一个词为词汇表中的任意词的概率都计算一遍。使用负采样样本作为所有非当前词 w(i)的替代。它的实现有多种算法，比如根据词频的带权采样算法。

3.3 模型训练

本文采用 Python 的 gensim 模块提供的 Word2Vec 工具包。训练过程中影响实验准确性和效率的参数主要如表 2 所示:

表 2 关键参数及解释

参数	解释
sg	训练模型选择，取 0 为 CBOW 模型；取 1 为 Skip-Gram 模型
hs	优化方法选择，取 0 为 NS 方法；取 1 为 HS 方法
negative	负采样样本值，默认为 5
size	词向量维度，一般而言，几十到几百之间效果比较好
min_count	词频最低值，一般为 10~100 之间，用于限制词汇量大小
sample	高频词采样样本数, Google 文档推荐值为 1e-5~1e-3 之间
window	训练窗口大小，表示句子中当前词和预测词最远距离，一般取值越大越好，直到某个临界值
workers	训练模型的并行线程数，一般取 4~6

sg 参数对应模型的选择，取 1 代表 Skip-Gram 模型；取 0 代表 CBOW 模型。hs 参数对应优化算法的选择，取 1 代表 Hierarchy SoftMax 算法；取 0 代表 Negative Sampling 算法。negative 参数对应 Negative Sampling 算法中负采样样本的数量。size 是词向量的维度，随着 size 值的增大，词向量准确性会先提高，到达某极值后，size 值继续增加，准确性反而会降低。min_count 参数是用来过滤低频词的，相当于进行一次词频低于 min_count 的词删除的预处理。sample 参数是对高频词进行处理的。迭代过程中更新高频词会占用一定的时间，而高频词对应的词向量变化不大，

①例如，句子“杭州绿茶真的太好喝了”，包含 4 个三元词组: “杭州绿茶真的”、“绿茶真的太”、“真的太好喝”、“太好喝了”；其实它的含义为“杭州绿茶好喝”，却没有一个词组表达了这个意思，如果允许跳过 2 个词，则会出现 18 种三元词组，其中一种为“杭州绿茶好喝”。

故采用 Subsampling 技术(二次采样)在训练时跳过某些高频词。如公式(10)所示, p(w)代表词语 w 被跳过的概率, 其中 f(w)为该词在语料库 C 中出现的概率: f(w)>t 时, f(w)越大, p(w)越大, 被跳过的概率越大。

$$p(w)=1-\sqrt{\frac{t}{f(w)}} \tag{10}$$

window 参数指训练窗口的大小, 与上下文构造相关。每次构造词 w 的上下文 context(w)时, 生成[1, window]上一个随机整数 c, 在 w 前后各取 c 个词, 构成 context(w)。一般而言, window 值越大越好, 直到到达某个极值。

workers 参数则是并行线程数, workers 越大, 训练速度越快; 可以根据计算机性能尽可能增加 workers 值。

3.4 疾病主题语义相似性计算

模型训练的结果是将每个疾病主题词都映射为 N 维向量空间中的一个点, 根据向量空间中余弦距离公式求解词向量之间的距离, 作为其语义相似性。假设两个疾病主题的 N 维词向量分别为:

$$\begin{aligned} t_1&=(w_{11},w_{12},w_{13},\cdots,w_{1(n-1)},w_{1n}),\\ t_2&=(w_{21},w_{22},w_{23},\cdots,w_{2(n-1)},w_{2n}) \end{aligned} \tag{11}$$

余弦值越大, 表示疾病主题 t_1 和疾病主题 t_2 在语义上越相似。计算公式如下:

$$\cos(\theta)=\frac{\sum_{k=1}^{n}w_{1k}w_{2k}}{\sqrt{\sum_{k=1}^{n}w_{1k}{}^2}\sqrt{\sum_{k=1}^{n}w_{2k}{}^2}} \tag{12}$$

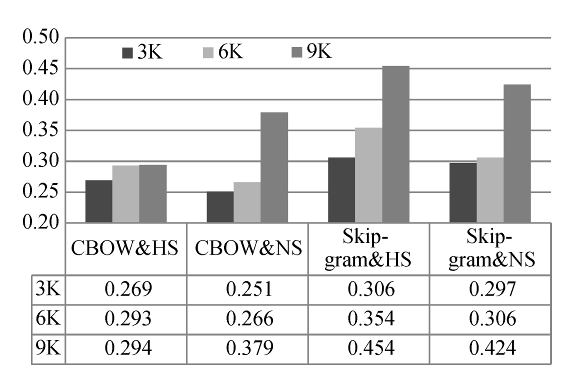
分别计算 30 个疾病主题两两之间的语义距离, 得到 435 组值。

4 实验设计与结果

Word2Vec 的效果受到数据规模、模型的选择、参数的设定等因素影响。实验将对上述内容进行一一检测, 并与专业医生对 30 种疾病关联关系的评分结果对比。记专家评分值为 base 值, 利用 SPSS 计算训练值与 base 值的相关性分析, 可得到各种因素对结果的影响, 并评估该方法在实际中的可用性。

4.1 数据规模

图 4 中的纵坐标代表训练结果与 base 值的 Pearson 相关系数。数据集从 3K 扩大到 6K 时, 效果有了一定的提高, 而扩大到 9K 时, 相关系数明显增大。另外在 3K 数据集下, 即使效果最好的 Skip-Gram&HS 结果, 相关关系仍然不太显著。分别对各类参数进行调整, 使用 3K 数据集时最优结果在 0.01 水平下相关系数为 0.394, 小于 0.4。而数据集增大到 9K 时, Skip-Gram 模型的初始相关系数就达到 0.454。可见数据规模是影响 Word2Vec 训练质量的关键性因素。数据规模越大, 模型效果越好。



	CBOW&HS	CBOW&NS	Skip-gram&HS	Skip-gram&NS
3K	0.269	0.251	0.306	0.297
6K	0.293	0.266	0.354	0.306
9K	0.294	0.379	0.454	0.424

图 4 不同数据规模下模型结果与专家打分的相关系数对比

Word2Vec 是基于词的上下文关系来建立词的语义关系, 数据集增大, 词的上下文语境更全, 训练得到的词向量更能够反映出该语料集里词汇的语义。实验说明 3K 数据集过小, 难以很好地衡量词与词之间的语义相似性。Word2Vec 技术在样本数较小时表现并不好。

4.2 模型选择

图 4 对 4 种训练架构的结果也进行了比较, Skip-Gram 效果明显比 CBOW 好; 但是后者的实际运行时间较短。对于同一语料库, Skip-Gram 会利用"跳过某些单元"的方式来扩大训练样本, 这也可以看成"数据规模"的增加, 从而带来了模型性能的增加与训练时间的增长。

从优化算法来看, 以 Skip-Gram 模型为前提, 虽然在数据集为 9K 时, Skip-Gram&HS 比 Skip-Gram&NS 准确一些, 但在 3K 和 6K 数据集下两者差别并不大, 见图 4。此处用于比较的 Negative Sampling 采样中负样本取值(negative)为 5, 事实上, 负样本取值也会影响 Negative Sampling 方法的效果。

表 3 显示了对负样本取值的进一步对比。由于 9K 数据集训练时间太长, 先缩小词向量维度 size 值为 50 以缩短训练时间, 再探究 negative 值对结果的影响。

在 Skip-Gram&NS 方法中，negative 值越大，相关系数越高。虽然 negative 取 5 时，训练结果不如 Hierarchy SoftMax 算法；但当 negative 取 20 时，相关系数达到 0.539，比 Hierarchy SoftMax 算法高很多。

表 3　对 negative 因素的对比(sg=1, size=50, 9K 数据集)

Skip-Gram 模型		Hierarchy SoftMax	Negative Sampling (negative 取值)		
			10	15	20
base	Pearson 相关性	.497**	.511**	.521**	.539**
	N	435	435	435	435

综上所示，Skip-Gram 和 Negative Sampling，当负样本采样值为 20 时，训练模型得到的结果较优。以下参数选择均以 Skip-Gram&NS 方法为前提来开展实验。

4.3　参数对比

为了掌握词向量维度的大小(size)对算法结果的影响，首先利用规模较小、训练速度较快的数据集寻找 size 参数变化对结果的影响，再选择可使结果达到最优的参数取值区间进一步观察。在 3K 数据集控制 size 值在[50, 500]范围内，发现词向量维度值并不是越大越好，取值在[50, 100]结果较优，如图 5 所示。进而在 9K 数据集缩小 size 的取值在[50, 100]区间范围，发现词向量维度为 50 的时候，Skip-Gram&NS 与专家评分的相关性最高，如图 6 所示。

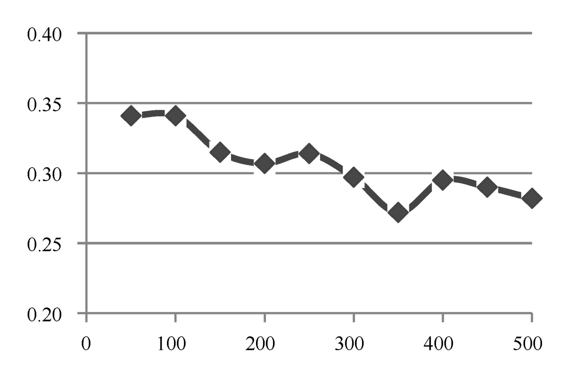


图 5　size 因素取值对结果影响(sg=1, 3K 数据集)

在 Skip-Gram&NS 方法中，固定已测参数负采样值、词向量维度为最佳取值(negative=20, size=50)，研究高频采样阈值 sample 对结果的影响。在 Google 给出的 Word2Vec 工具包中推荐在[1e-5, 1e-3]范围内改变 sample 值，因此将该参数设置为图 7 所示的几个取

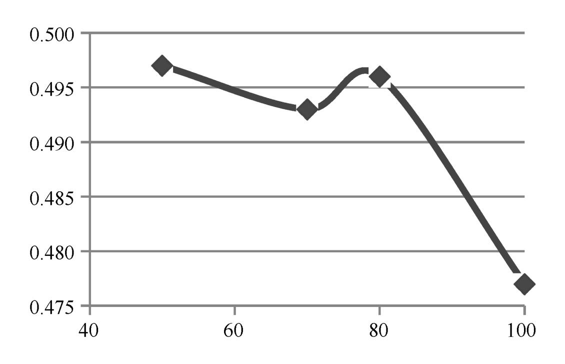


图 6　在图 5 的最优区间细粒度观察 size 取值对结果的影响(sg=1, hs=1, 9K 数据集)

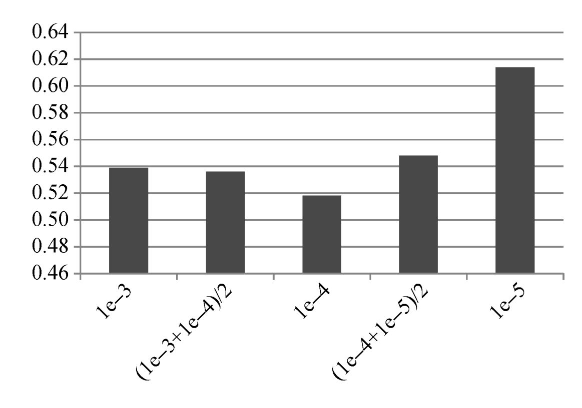


图 7　sample 因素对结果的影响(sg=1, hs=0, negative=20, size=50, 9K 数据集)

值。结果发现，sample 值越小训练结果与 base 值的相关性越高，训练时间也明显变短。高频词在语料库中出现次数很多并且提供的有用信息更少，训练时对应的词向量变化也较小。由公式(10)可知，sample 值越小，语料库中出现概率高于 sample 的词语越多，二次采样中，被跳过的高频词越多，准确性越高。在 9K 数据文本作为训练集的情况下，sample 值取 le-5 时结果较优。两者相关系数达到 0.614，结果有明显提高。

令 sample=1e-5，进一步考察低频词阈值 min_count 对其取值从 40 开始，以步长为 20 变化。由表 4 可知，在(40, 100)范围内 min_count 的改变对结果影响不大。训练前创建词表时，去掉词频低于 40 的词可以使结果更优。

表 4　min_count 因素对结果的影响(sg=1, hs=0, negative=20, size=50, sample=1e-5, 9K 数据集)

min_count 参数		40	60	80	100
base	Pearson 相关性	.614**	.610**	.611**	.606**
	N	435	435	435	435

上下文窗口 window 的取值理论上是越大越好，但是 window 扩大将致使训练时间加长。在[50, 200]区间上改变 window 参数，与专家评分比较得到如图 8 所示的结果：window 为 50 左右相关系数就不再增加，为 0.635。此时 Skip-Gram 取上下文样本时，在[1, 50]区间上生成一个随机整数 c，然后在词 w 前后各取 c 个词，构成 context(w)。

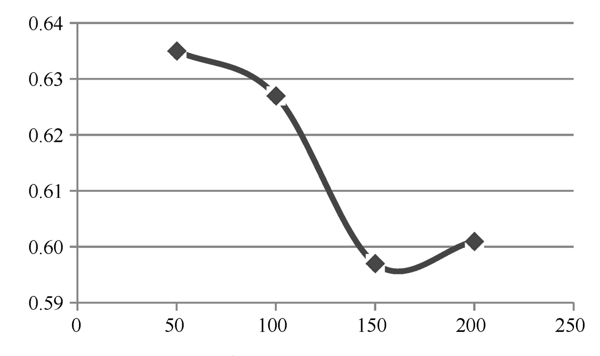


图 8　window 因素对比(sg=1, hs=0, negative=20, size=50, sample=1e-5, min_count=40, 9K 数据集)

综合以上对参数因素的探究，发现采用 Skip-Gram 模型和 Negative Sampling(负样本采样值为 20)算法组合，词向量维度取 50，高频词采样阈值取 1e-5，低频词阈值取 40，上下文窗口取 50 的时候，训练模型得到的相似性度量结果最优，与 base 值的相关系数达到 0.635，将此结果记为 W2V。

5　Word2Vec 的疾病关联探测效果分析

将实验得到的最优结果 W2V 与 base 值进行详细对比分析，对 W2V 值归一化处理，得到散点图如图 9 所示。将 435 组值按照 base 值从大到小排序，从 1 开始编号到 435，以编号作为横坐标，base 值和 W2V 值作为纵坐标。

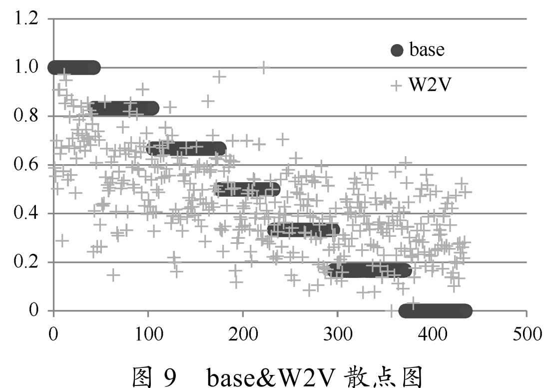


图 9　base&W2V 散点图

在 base 值从大到小降低时，W2V 值也是整体降低的趋势。语义相似性较高的区域，W2V 值更分散；相似性低的区域，W2V 的值相对而言更集中。Word2Vec 计算结果的整体性较好，局部性需要进一步改进。

按照 base 值的取值范围，将数据分成 7 个区间，对应的 base 值从高到低。由表 5 容易发现，base 值较高的区间上，相应的 W2V 的最小值要比低区间上最小值高。均值的分布与 base 值区间值的高低变化趋势一致，区间越高，均值越大，这也与图 9 结论相吻合；但是相对而言，W2V 的均值变化范围更小。中位数的变化趋势与均值基本保持一致。base 取 1 的区间上，标准差最大，为 0.157；base 取 0 的区间上，标准差最小，为 0.127；这表示，相似性最高的区间上，W2V 点的离散程度更大，相似度最低的区间上，W2V 点的分布情况最集中。

表 5　W2V 值在不同水平 base 值上统计值

base 值	编号范围	W2V 值				
		最小值	最大值	均值	中位数	标准差
1	2:43	0.244	0.971	0.709	0.710	0.157
0.833	44:105	0.148	0.910	0.562	0.562	0.157
0.667	106:176	0.162	0.962	0.522	0.522	0.150
0.5	177:233	0.119	1.000	0.426	0.399	0.156
0.333	234:296	0.085	0.704	0.372	0.384	0.147
0.167	297:372	0.000	0.572	0.327	0.338	0.130
0	373:436	0.032	0.609	0.311	0.282	0.127

由 Word2Vec 训练得到的疾病主题语义关联中，按照相似性从高到低排序，得到的前 10 对相似的疾病主题对如表 6 所示。对应的 base 值中，有 6 组疾病是高度相关的，3 组疾病相关性也较高。唯有男性健康与女性健康这组关系，Word2Vec 计算得到是高度相关，而专家评分仅为 0.5，差异较大。这可能是因为 Word2Vec 中表示词组时用向量相加表示，Men's Health、Women's Health 的向量分别为词语 health 与词语 men、women 的向量加合，计算时两者相似性会随之增高。尤其词语 men 与词语 women 在语料库中出现频率较高，特指性不强，在语义上还很相似，从而高估了 Men's Health(男性健康)与 Women's Health(女性健康)之间的关联。并且这两种疾病本身涵盖的范围也比较宽泛，影响了 Word2Vec 判断准确性。

表 6 Top10 相似疾病主题

疾病主题 i	疾病主题 j	W2V 相似性	W2V 归一化	对应 base 值
Men's Health	Women's Health	0.848	1.000	0.5
Cardiovascular	Cholesterol	0.822	0.971	1
Breast Cancer	Prostate	0.814	0.962	0.667
Cardiovascular	Heart Disease	0.801	0.948	1
Mental Health	Women's Health	0.767	0.910	0.833
Anxiety	Depression	0.766	0.909	1
Allergy	Asthma	0.766	0.908	1
Cardiovascular	Hypertension	0.746	0.886	1
Cardiovascular	Stroke	0.729	0.867	1
Men's Health	Mental Health	0.724	0.862	0.667

6 结 语

本文选取 30 个疾病主题，采集 Medical News Today 上的新闻文本，利用 Word2Vec 技术计算疾病间关联关系，并与专家评分结果进行对比。研究发现，数据规模越大，模型效果越好，但训练时间更长；Skip-Gram 模型结合负样本数为 20 的 Negative Sampling 优化方法在大规模数据集上的实验结果最优；高频词二次采样阈值越小，训练效果越好，训练时间也越短。最优条件下，训练结果与专家评分的相关系数达到 0.635；语义相似性较高的区域，Word2Vec 训练值更分散；相似性低的区域，Word2Vec 训练值相对而言更集中。将 Word2Vec 训练结果按照相似性从高到低排序，得到的前 10 组疾病关系中，有 9 组在专家评分中相关性也很高。

利用 Word2Vec 技术在面向大众的健康信息源中也可以探测疾病关联，其有效性表明该技术可用于改善大众的健康信息搜寻的个性化服务。

未来将从以下方面开展研究：扩大数据集，Word2Vec 在数据规模增大时效果提升明显，实际中使用更多数据可得到更理想的结果；调整疾病类型，从更细的粒度开展关联关系研究。

参考文献：

[1] Kempson E. Review Article: Consumer Health Information Services [J]. Health Libraries Review, 1984, 1(3): 127-144.

[2] Eysenbach G. Recent Advances: Consumer Health Informatics [J]. BMJ Clinical Research, 2000, 320(7251): 1713-1716.

[3] 侯小妮，孙静．北京市三甲医院门诊患者互联网健康信息查寻行为研究[J]．图书情报工作，2015，59(20)：126-131，11. (Hou Xiaoni, Sun Jing. Research on Internet Health Information Searching Behaviors of Outpatients from Tertiary Referral Hospital in Beijing [J]. Library and Information Service, 2015, 59(20): 126-131, 11.)

[4] Klavans J L, Muresan S. Evaluation of the DEFINDER System for Fully Automatic Glossary Construction[C]. In: Proceedings AMIA Annual Symposium. 2001: 324-328.

[5] Zeng-Treitler Q, Tse T. Exploring and Developing Consumer Health Vocabularies [J]. Journal of the American Medical Informatics Association, 2006, 13(1): 24-29.

[6] Zeng-Treitler Q, Goryachev S, Tse T, et al. Estimating Consumer Familiarity with Health Terminology: A Context-based Approach [J]. Journal of the American Medical Informatics Association, 2008, 15(3): 349-356.

[7] Burgun A, Bodenreider O. Mapping the UMLS Semantic Network into General Ontologies [C]. In: Proceedings of Annual Symposium. 2001: 81-85.

[8] Keselman A, Smith C A, Divita G, et al. Consumer Health Concepts that do not Map to the UMLS: Where do They Fit? [J]. Journal of the American Medical Informatics Association, 2008, 15(4): 496-505.

[9] Yang Z H, Lin H F, Li Y P, et al. TREC 2005 Genomics Track Experiments at DUTAI [C]. In: Proceedings of the 14th Text REtrieval Conference. 2005: 1-9.

[10] Yang Z H, Lin H F, Li Y P, et al. DUTIR at TREC 2006 Genomics and Enterprise Tracks [C]. In: Proceedings of the 15th Text REtrieval Conference. 2006: 1-10.

[11] Jiang Q, Wang Y, Hao Y, et al. miR2Disease: A Manually Curated Database for microRNA Deregulation in Human Disease [J]. Nucleic Acids Research, 2009, 37(Database issue): D98-104.

[12] Yang H, Yang C C. Using Health Consumer Contributed Data to Detect Adverse Drug Reactions by Association Mining with Temporal Analysis [J]. ACM Transactions on Intelligent Systems & Technology, 2015, 6(4): Article No.55.

[13] Chen A T. Exploring Online Support Spaces: Using Cluster Analysis to Examine Breast Cancer, Diabetes and Fibromyalgia Support Groups [J]. Patient Education and Counseling, 2012, 87(2): 250-257.

[14] 刘红霞，张进，陈璟浩．WHO 英文网站健康主题语义链接关系社会网络分析[J]．图书情报工作，2014，58(13)：75-82. (Liu Hongxia, Zhang Jin, Chen Jinghao. Social Network

Analysis of Semantic Links Relationships Among Health Topics in WHO English Website [J]. Library and Information Service, 2014, 58(13): 75-82.)

[15] Bengio Y, Schwenk H, Senécal J-S, et al. A Neural Probabilistic Language Model [J]. Journal of Machine Learning Research, 2003, 3(6): 1137-1155.

[16] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space [OL]. [2016-05-13]. http://arxiv.org/pdf/1301.3781v3.pdf.

[17] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and Their Compositionality [A]. //Advances in Neural Information Processing Systems [M]. 2013: 3111-3119.

[18] Handler A. An Empirical Study of Semantic Similarity in WordNet and Word2Vec [D]. Columbia University, 2014.

[19] Amunategui M, Markwell T, Rozenfeld Y. Prediction Using Note Text: Synthetic Feature Creation with Word2Vec [J]. Computer Science, 2015(3): 1-6.

[20] Ju R, Zhou P, Li C H, et al. An Efficient Method for Document Categorization Based on Word2Vec and Latent Semantic Analysis [C]. In: Proceedings of the 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM). IEEE, 2015: 2276-2283.

[21] Su Z, Xu H, Zhang D, et al. Chinese Sentiment Classification Using a Neural Network Tool — Word2Vec [C]. In: Proceedings of the 2014 International Conference on Multisensor Fusion and Information Integration for Intelligent Systems (MFI). IEEE, 2014: 1-6.

作者贡献声明:

陈翀: 提出研究思路, 设计研究方案, 论文修订;
罗文馨: 进行实验, 采集、清洗和分析数据, 论文起草;
邓思艺: 参与文献调研。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: luowenxin1994@foxmail.com, chenchong@bnu.edu.cn。

[1] 罗文馨. Word2Vec 疾病相似性结果.xlsx. 本文中与专家评分结果相关度最高的(Pearson 0.635)Word2Vec 结果(即 W2V)得到的疾病关联度.

[2] 罗文馨. Word2Vec 词向量训练结果.txt. W2V 结果对应的词向量, 用于计算疾病的语义相似性.

[3] 陈翀. 30 种疾病新闻网页.html. 抓取的 Medical News Today 网站上的 9 000 个新闻网页.

收稿日期: 2016-05-16
收修改稿日期: 2016-05-22

Detecting Disease Associations with Word2Vec from Consumer Health Information

Luo Wenxin　Chen Chong　Deng Siyi
(School of Government, Beijing Normal University, Beijing 100875, China)

Abstract: [**Objective**] Average people usually do not know the complex associations among diseases, which poses negative effects to their health information seeking experience. This study tries to detect the associations among diseases using popular medical information with the help of deep learning technology (Word2Vec), aiming to improve personalized information services. [**Methods**] First, we identified 30 common disease topics with the help of medical professionals, and then collected related reports from Medical News Today. Second, we built word vector for each document with Word2Vec technology to calculate the semantic similarities among them. Finally, we compared the machine training results with experts' scores to evaluate the performance of the proposed method. We also investigated the impacts of different models, optimization methods, data sizes and important parameters to the results. [**Results**] The correlation coefficient between the Word2Vec results and the experts' scores reached 0.635 in optimal condition. We found that Skip-Gram model with less than 20 negative samples on large scale dataset yielded the best results. [**Limitations**] The precision of the Word2Vec judgment was affected by the number of disease topics. The granularity of disease topic needed to be improved. [**Conclusions**] The Word2Vec technology could be used to identify diseases association from consumer health information sources. It could also be used to improve the personalized health information services.

Keywords: Word2Vec　Disease association　Non-professional medical information　Health informaiton　Personalization

---

EBSCO 进一步资助 Koha

EBSCO 宣布继续倡导开源和开放获取，进一步资助 Koha。Koha 是世界上第一个功能丰富、免费开源的集成图书馆系统，全世界有 15 000 多家各种类型的图书馆使用 Koha 作为其集成图书馆系统。

EBSCO 于 2015 年 2 月起开始为 Koha 提供资金资助，本次对 Koha 的最新资助，将帮助 Koha 进行下一阶段的功能改进，如额外的系统互操作性，以及采购和电子资源管理功能，具体来说，包括:

(1) 开发一个采购的 API;

(2) 全面实施订购和发票系统;

(3) 改进 Koha 和 CORAL 的互操作性，为传统的集成图书馆系统工作流和 ERM 功能的结合提供一个开源的解决方案。

Koha 将坚持开源的传统，这些 Koha 的增强功能也将是开源的，可供他人使用、修改和重新部署。这些增强的功能可望于 2017 年第一季度完成。

(编译自: https://www.ebsco.com/news-center/press-releases/ebsco-information-services-continues-to-support-open-source-technology)

(本刊讯)